Characterizing the public perception of WhatsApp through the lens of media

Josemar Alves Caetano[1], Gabriel Magno[1], Evandro Cunha[1,2],Wagner Meira Jr.[1],
Humberto T. Marques-Neto[3], Virgilio Almeida[1,4]
{josemarcaetano, magno, evandrocunha, meira}@dcc.ufmg.br,
humberto@pucminas.br, virgilio@dcc.ufmg.br

[1] Dept. of Computer Science, Universidade Federal de Minas Gerais (UFMG), Brazil
[2] Leiden University Centre for Linguistics (LUCL), The Netherlands
[3] Dept. of Computer Science, Pontifícia Universidade Católica de Minas Gerais (PUC Minas), Brazil
[4] Berkman Klein Center for Internet & Society, Harvard University, USA

CEUR-WS.org/Vol-2482/paper42.pdf

Abstract

WhatsApp is, as of 2018, a significant component of the global information and communication infrastructure, especially in developing countries. However, probably due to its strong end-to-end encryption, WhatsApp became an attractive place for the dissemination of misinformation, extremism and other forms of undesirable behavior. In this paper, we investigate the public perception of WhatsApp through the lens of media. We analyze two large datasets of news and show the kind of content that is being associated with WhatsApp in different regions of the world and over time. Our analyses include the examination of named entities, general vocabulary and topics addressed in news articles that mention WhatsApp, as well as the polarity of these texts. Among other results, we demonstrate that the vocabulary and topics around the term "whatsapp" in the media have been changing over the years and in 2018 concentrate on matters related to misinformation, politics and criminal scams. More generally, our findings are useful to understand the impact that tools like WhatsApp play in the contemporary society and how they are seen by the communities themselves.

Copyright © CIKM 2018 for the individual papers by the papers' authors. Copyright © CIKM 2018 for the volume as a collection by its editors. This volume and its papers are published under the Creative Commons License Attribution 4.0 International (CC BY 4.0).

1 Introduction

The messaging service WhatsApp is, as of 2018, one of the most rapidly growing components of the global information and communication infrastructure, counting with 1.5 billion users who send around 60 billion messages per day [Con18]. This tool combines one-to-one, one-to-many and group communication by offering private chats, broadcasts and public group chats, through which users are able to send text and media (audio, image and video), as well as files in various formats.

According to data published by Statista [Sta18], more than half of the population of Saudi Arabia, Malaysia, Germany, Brazil, Mexico and Turkey were active WhatsApp users in 2017. Also, the Reuters Institute Digital News Report 2018 [NFK+18] shows a rise in the use of messaging applications, including WhatsApp, as sources of news in several parts of the world. This report indicates that WhatsApp use for news has almost tripled since 2014 and it has surpassed Twitter as a communication system in many countries. One of the alleged reasons for this is that users are looking for more private and secure spaces to communicate. In addition to this, WhatsApp turned out to be an important platform for political propaganda and election campaigns, having held a central role in elections in Brazil, India [Goe18], Kenya, Malaysia, Mexico and Zimbabwe, for instance. Also, WhatsApp has been frequently associated with the spread of misinformation and disinformation [Wat18].

Despite its prominence, continued growth and opacity, there has been an insufficient number of studies exploring the various aspects of WhatsApp and similar mobile messaging applications [GWCG18]. Since WhatsApp provides encrypted end-to-end communi-

cation, it is a great challenge to conduct large-scale analyses on the behavior of its users. In this work, we take a different approach: instead of looking at inside the system, we focus on the public perception of WhatsApp from outside sources. The goals of this paper are:

- to characterize how media in different countries interpret the role of WhatsApp in society;
- to analyze the evolution of the perception of WhatsApp over time, from its creation until its massive popularization;
- to comprehend how sensitive topics, such as politics, crime and extremism, are related to WhatsApp in different regions of the world and in distinct periods of time.

To achieve these goals, we explore different techniques: analysis of Web search behavior, co-occurring named entities and vocabulary, co-occurrence networks, topics addressed and textual polarity. According to our understanding, each of these methods is able to provide additional information about the perception of WhatsApp in the news articles investigated. As a whole, our results indicate that the media has significantly changed its perception and portrayal of WhatsApp: while in the period before 2013 the focus of the news was on WhatsApp features, in the following years the tool started to be more associated with social issues, including the dissemination of misinformation.

This paper is organized as follows: in Section 2, we review a selection of works on WhatsApp and, more generally, on the use of textual datasets to understand social phenomena; in Section 3, we describe our methodology of data collection and the overall characterization of the datasets used in this investigation; next, in Section 4, we characterize the vocabulary, analyze the topics addressed and evaluate the polarity of the news articles contained in our datasets; finally, in Section 5, we conclude the paper and present future directions of work.

2 Related Work

On the use of textual datasets to understand social phenomena

Analyzing how a term is used over time and in a geographic location is important to help in the understanding of how cultural values, societal issues and customs are perceived by society and expressed through language [Cam13, Mat53]. *Culturomics*, for example, is a concept proposed by [MSA+11] referring to a method for the study of human behavior and cultural trends through quantitative analyses of texts, using sources like large collections of digitized books. Several studies explore this method to investigate topics such as the dynamics of birth and death of words [PTHS12], semantic change [GB11], emotions in literary texts [ALGB13] and characteristics of modern societies [Rot14]. Some works propose a complementary approach to *culturomics* by using historical news data [Lee11], analyzing European news media [FTA+10] or the writing style and gender bias of particular topics in large corpora of news articles [FALW+13]. Other works concentrate in specific events in history, such as the Fukushima nuclear disaster [LWSVC14], by using large datasets of media reports to understand aspects such as how the media polarity towards a topic changes over time.

Employing methods similar to the ones presented here, [CMC+18] investigate the perception and the conceptualization of the term "fake news" in the media, showing that contextual changes around this expression might be observed after the United States presidential election of 2016. However, as far as we are concerned, this is the first work that uses these methods to examine in detail how the term "whatsapp" is being reported by news media in different parts of the world, making us able to analyze how important topics, such as misinformation, manipulation and extremism, might be associated with WhatsApp by societies.

On WhatsApp

Despite the increasing use of WhatsApp in the world, few quantitative and large-scale studies about this instant messaging application are currently available. [GT18] propose a data collection methodology for this application and perform a statistical exploration to indicate how data from WhatsApp public groups can be collected and analyzed. Also, [MGB17] collect WhatsApp messages to monitor critical events during Ghana's 2016 presidential election, and [CdO13] analyze differences between WhatsApp and SMS messaging system using a large-scale survey. [FCSD15] investigate Facebook and WhatsApp traces collected from an European national wide mobile network and characterize the usage of both applications. The work of [SHS+16] surveys users to investigate the usage of WhatsApp groups and, more specifically, its implications for mobile network traffic, while [RSS+18] collect personal information and messages from one hundred WhatsApp users with the aim of understanding their usage patterns.

All of these works investigate a limited part of WhatsApp, therefore offering a restricted understanding of how this application is used. Nevertheless, here we study this tool using large datasets of external

data provided by news articles containing the term "whatsapp" in different regions of the world and covering the whole WhatsApp history, thus shedding light not exactly on its usage, but on how it is viewed from outside sources.

3 Data Collection

We use two large datasets of news articles in this study. The first one is a collection of texts from the Corpus of News on the Web (NOW Corpus), which contains articles from online newspapers and magazines written in English in 20 different countries from 2010 to the present time [Dav13]. This corpus is available for download and online exploration[1] and, according to its author, it is, at the moment of our data collection, the largest corpus available in full-text format. In 31 May 2018, we gathered all the news articles containing the 33,185 occurrences of the term "whatsapp" in the NOW Corpus. These news articles cover every year in the corpus (from 2010 to 2018) and comprise all 20 countries represented. These countries were then grouped into six regions based on their geographic locations (Africa, British Isles, Indian subcontinent, Oceania, Southeast Asia and the Americas).

Our second dataset includes articles collected from Brazilian online newspapers and magazines, all written in Portuguese, also containing the term "whatsapp". We searched for articles starting from 2010, but did not find any from 2010 and 2011 containing the term "whatsapp", so our second dataset contains news from 2012 to 2018. To build this dataset, we used the tool `Selenium`[2] to automate Web searches with the term "whatsapp" in the following ten major Brazilian news websites: *Exame*, *Folha de S. Paulo*, *Gazeta do Povo*, *G1*, *O Estado de S. Paulo*, *R7*, *Terra*, *Universo Online* (*UOL*), *Valor Econômico* and *Veja*. The total number of occurrences of "whatsapp" extracted from these websites on 31 May 2018 is 4,047. Finally, we used the Python library `newspaper`[3] to collect the full texts of these news articles.

In Sections 4.2 to 4.6, we analyze the news texts from the two previously described datasets. Table 1 shows the number of news containing the term "whatsapp" in our two datasets, according to the geographical origin of the corresponding news media and the year of publication of the news article.

In addition to these datasets, we also collected data from Google Trends[4], an online tool that indicates the frequency of particular terms in the total volume of searches in the Google Search engine. This tool also indicates the most common associated terms and the countries from which the highest volume of searches are originated from. It is also possible to filter these results for given periods. For our investigations, we collected data from searches made between 2010 and 2018, and use this information in Section 4.1.

[1] `https://corpus.byu.edu/now/`
[2] `https://www.seleniumhq.org/`
[3] `https://pypi.org/project/newspaper/`
[4] `https://trends.google.com/trends/`

4 Analyses and Results

In this section, we discuss the outcomes of different analyses aimed to understand the perception of WhatsApp in the media. Each characterization is introduced by a description of how it may contribute to accomplish our goals, followed by the methodology employed and, finally, by a presentation and discussion of the results found.

4.1 Web search behavior

Before analyzing the public perception of WhatsApp through the lens of news articles from different regions of the world, we investigate whether it is possible to observe a change in the Web search behavior regarding the term "whatsapp" through time. We use data collected from Google Trends to perform this analysis.

Our results show that, unsurprisingly, the number of queries on the Google Search engine for the term "whatsapp" is constantly growing since the release of this tool for Android devices in 2010, as indicated in Figure 1. Also, Table 2 lists the five most frequent search terms employed by users who also searched for "whatsapp" from 2010 to 2018. Here, we notice a shift in the related terms through the years: in the first two years, most of the words are concerned with the download of the app ("download", "descargar") and device compatibility ("blackberry", "iphone", "nokia"); then, from 2012 onwards, queries for "whatsapp" start to be linked to different topics, especially features of the tool ("status unavailable", "whatsapp encryption", "video status download"), but also content shared in WhatsApp ("imagens para whatsapp", "el negro del whatsapp").

4.2 Co-occurring named entities

In natural language processing, *named entity recognition* is the task of extracting mentions of named entities – that is, definite noun phrases referring to individuals, organizations, dates, locations – in a text [BLK09]. We here extract the most mentioned named entities in our NOW Corpus dataset for each region and year of publication of the articles in order to understand who are the main actors related to the tool WhatsApp according to the media. In this paper, the co-occurrence is computed on a document level, so we consider all the entities that are mentioned in

Table 1: (a) Number of news articles containing the term "whatsapp" in our NOW Corpus dataset according to the geographical origin of the corresponding news media; (b) Number of news articles containing the term "whatsapp" in both NOW Corpus and Brazilian news articles datasets according to the year of publication.

(a) Geographical origin of news articles in our NOW Corpus dataset

Region	**Country**	**Occurrences**
The Americas	United States	1,244
	Canada	507
	Jamaica	151
		Total: 5.73% / 1,902
Southeast Asia	Singapore	2,889
	Malaysia	2,578
	Philippines	253
	Hong Kong	124
		Total: 17.61% / 5,844
British Isles	Great Britain	2,251
	Ireland	2,152
		Total: 13.27% / 4,403
Africa	South Africa	5,274
	Nigeria	1,607
	Kenya	1,585
	Ghana	754
	Tanzania	3
		Total: 27.79% / 9,223
Oceania	Australia	895
	New Zealand	306
		Total: 3.62% / 1,201
Indian subcontinent	India	8,991
	Pakistan	1,353
	Sri Lanka	186
	Bangladesh	82
		Total: 31.98% / 10,612

(b) Year of publication of news articles in both NOW Corpus and Brazilian news articles datasets

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	**Total**
Occurrences in NOW Corpus	4	41	145	393	1,101	1,642	7,266	11,677	14,636	33,185
Occurrences in Brazilian news articles	0	0	4	91	427	785	904	888	948	4,047

our news articles as co-occurring with the key-term "whatsapp".

To perform the named entity recognition, we use the Natural Language Toolkit (NLTK)[5] classifier trained to recognize named entities. Since this tool does not support texts in Portuguese, we do not include the dataset containing the Brazilian news articles in this analysis.

Table 3 lists the ten most mentioned entities in each different region considered in this investigation. Overall, we observe that the most mentioned entities accompanying the term "whatsapp" are usually other social media companies ("Facebook", "Twitter"), countries ("US", "India"), cities ("Dublin", "Delhi") and demonyms ("African", "Australian"). When we analyze the continuation of the lists (not displayed here due to space constraints), we also find that US-American individuals like Mark Zuckerberg and Donald Trump are highly mentioned across the globe. However, local entities are also mentioned in their respective regions: among the entities not displayed in the table, the most mentioned persons or organized groups in each region are Mark Zuckerberg

[5] http://www.nltk.org/

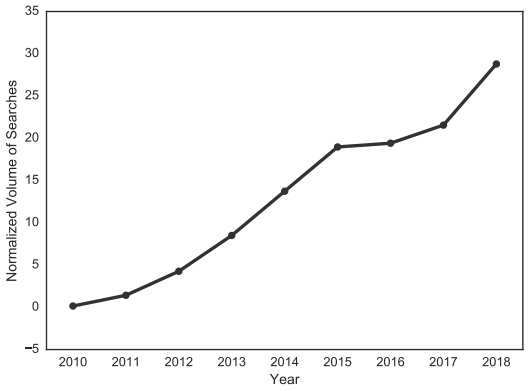


Figure 1: Normalized volume of queries for the term "whatsapp" in Google Search from 2010 to 2018

Table 2: Most frequent queries related to "whatsapp" on Google Search per year

Year	Search terms
2010	blackberry, iphone, service, download, android
2011	blackberry, nokia, app, download, descargar
2012	error status, status unavailable, zello, sniffer download, double check
2013	ios 7, imagens para whatsapp, intell app up, pagare whatsapp, baixa whatsapp
2014	blaue haken, for nokia xl, masti.com, facebook compra whatsapp, blue ticks on whatsapp
2015	whatsapp web, whatsapp reborn, caling feature, whatsapp transparante, llamadas whatsapp
2016	negrita whatsapp, gb whatsapp, whatsapp encryption, el negro del whatsapp, cartas y whatsapp
2017	video status download, whatsapp plus 2017, status tamil, wasap weed, whatsapp storing
2018	gb whatsapp 2018, plus 2018, call girls group link, browserling, whatsapp business

(the Americas and Oceania), Barisan Nasional (Southeast Asia), Paddy Jackson (British Isles), Uhuru Kenyatta (Africa) and Narendra Modi (Indian subcontinent). These findings suggest that news regarding the WhatsApp tool might deal with locally relevant entities – which also are, most of the times, related to the local political scenarios.

The ten most mentioned entities in each year are displayed in Table 4. Among the entities that do not appear in the table due to space limitations, the most mentioned persons or organized groups in each year are: Steve Jobs (2011), Neil Papworth (2012), Mark Zuckerberg (2013 and 2014), Islamic State (2015 and 2016) and the Bharatiya Janata Party – BJP (2017 and 2018). This indicates that, in general, the most relevant entities in the articles ceased to be linked to technology (Jobs, Papworth, Zuckerberg) and started

Table 3: Most mentioned named entities in each region (the entity "whatsapp" is excluded from the lists)

Region	Entities
The Americas	Facebook, Google, US, Twitter, Instagram, Apple, Android, American, Europe, China
Southeast Asia	Facebook, Malaysia, India, Singapore, US, Malaysian, Google, Indian, China, Chinese
British Isles	Facebook, Ireland, US, London, Irish, Google, British, Android, Dublin, Twitter
Africa	Facebook, Twitter, South Africa, Nigeria, African, Kenya, Instagram, Africa, Nigerian, US
Oceania	Facebook, US, Australia, Google, Australian, Apple, Instagram, Twitter, Facebook Messenger, New Zealand
Indian subcontinent	India, Facebook, Indian, Delhi, Mumbai, Pakistan, BJP, US, Twitter, Google

to be related to social and political situations (Islamic State and BJP) from 2015 onwards, showing that WhatsApp gained importance outside of the world of technology and business.

4.3 Semantic fields of the surrounding vocabulary

Besides the analysis of the named entities that appear in the same news articles as the term "whatsapp", the investigation of the general vocabulary co-occurring with it is also valuable. One of the possible methods of performing such analysis is by observing the semantic fields (i.e. groups to which semantically related items belong) of the words that appear in our news articles datasets, so to detect relevant concepts mentioned in the texts [CMG+14]. Here, we use the tool `Empath`[6] [FCB16], which provides a set of 194 lexical categories representing different semantic fields, each containing a list of words. Since Empath is available only in English, the dataset containing Brazilian articles was again not included in this analysis.

For this task, we first extracted all the words of each article and applied lemmatization – that is, we grouped together their inflected forms so that they could be analyzed as single items based on their dictionary forms (*lemmas*). Lemmatization was performed employing the WordNet Lemmatizer function provided by the Natural Language Toolkit and using verb as the part-of-speech argument for the lemmatization method, as in [CMC+18]. Then, we counted the number of lem-

[6] `https://github.com/Ejhfast/empath-client`

Table 4: Most mentioned named entities in each year (the entity "whatsapp" is excluded from the lists)

Year	Entities
2010	Android, BlackBerry Messenger, BlackBerry, Kik, iPhone, Nokia, WiFi, Nokia N8, Symbian, India
2011	BlackBerry, iPhone, Facebook, Android, Skype, SMS, Google, Nokia, Apple, US
2012	Facebook, SMS, Android, India, iPhone, BlackBerry, US, Nokia, Skype, Twitter
2013	Facebook, Android, Twitter, Google, Apple, India, Skype, SMS, BlackBerry, Indian
2014	Facebook, Google, Twitter, US, Android, India, Apple, WeChat, China, Instagram
2015	Facebook, India, Twitter, Google, US, South Africa, Android, Instagram, Skype, Apple
2016	Facebook, India, Twitter, US, Google, Indian, Android, Instagram, Apple, iPhone
2017	Facebook, India, Twitter, US, Indian, Instagram, Google, London, South Africa, China
2018	Facebook, India, Twitter, US, Indian, Google, Instagram, Delhi, South African, Telegram

matized words that appeared in each one of the Empath categories. In this phase, instead of using the absolute frequency of words, we normalized it by dividing the frequency of words in each category by the total number of categorized words.

Since analyzing all the 194 Empath categories is impractical, we manually selected three relevant and noteworthy categories to scrutinize: *crime*, *government* and *law*. In Figure 2, we present the average proportion of words belonging to these categories in news articles representing different regions across the years. On the whole, we observe an overall increase in the proportion of words belonging to the three analyzed categories, with most of the peaks (such as the ones of 2013 in Oceania) probably due to political events (e.g. Australian federal election of 2013). This finding indicates that words from the semantic fields *crime*, *government* and *law* are being gradually more associated with WhatsApp in news from different regions of the world, corroborating the finding of Section 4.2 that shows an increase in the association of WhatsApp with social and political situations in recent years.

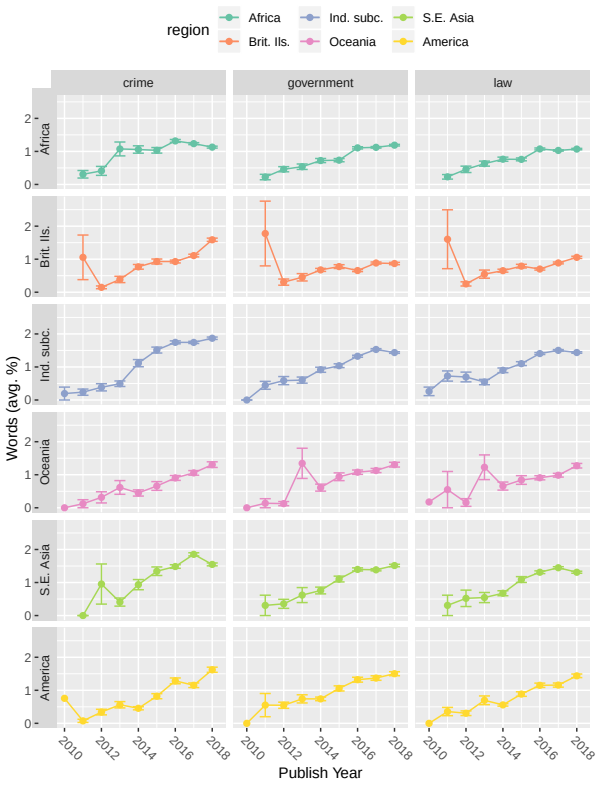


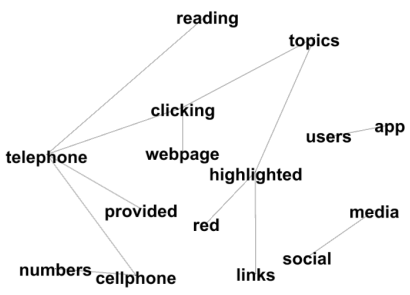
Figure 2: Average percentage of use of words from the semantic fields *crime*, *government* and *law* in different regions and years (bars indicate standard errors of the mean values)

4.4 Co-occurrence networks

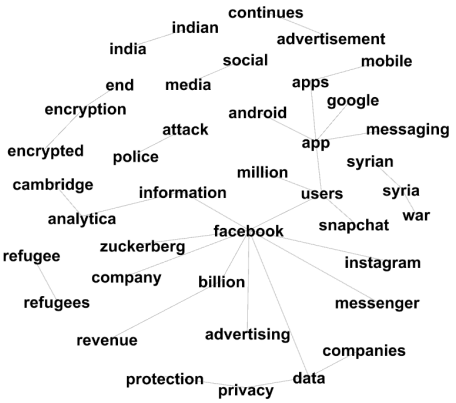
Another possible analysis on the vocabulary accompanying a key-term in a corpus can be made through the observation of *co-occurrence networks*. In our case, this method enables the visualization of the most relevant words that appear in the same news articles as the term "whatsapp" through the means of graphs. In this section, we consider both NOW Corpus and the Brazilian news articles dataset.

For this analysis, we first extracted all the words from the articles and removed stop words using the lists provided by the Natural Language Toolkit for English and Portuguese. Then, we extracted the most relevant words from each document by using the *term frequency-inverse document frequency* (tf-idf) technique, that reflects how important a word is to a document in a corpus [RU11]. We calculated the tf-idf for each pair (*document*, *word*) and extracted from the document the top 50 words with the highest tf-idf scores.

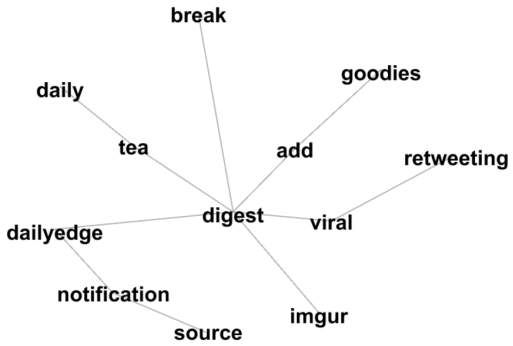
In the following step, we counted the number of co-occurrences of the pairs of words. For each document, we obtained the list with its 50 most relevant words (according to tf-idf) and incremented by one the



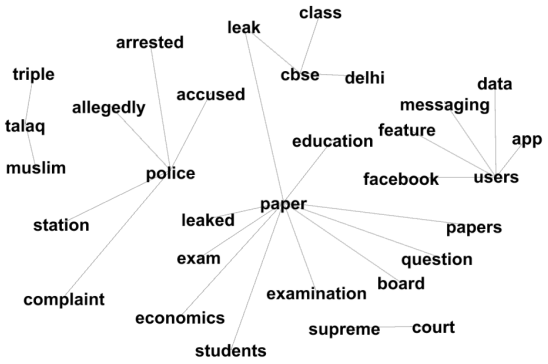
(a) Africa



(b) The Americas

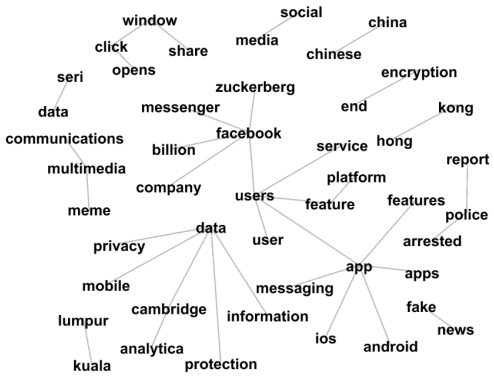


(c) British Isles



(d) Indian subcontinent

(e) Oceania



(f) Southeast Asia

Figure 3: Co-occurrence networks for NOW Corpus news articles

counter relative to each pair of words in this list (combination two by two). Instead of using the absolute count of articles in which two words co-occur, we normalized this value by dividing it by the total number of articles. At the end of this process, we obtained a graph in which vertices represent words and edges indicate their co-occurrence in the same texts.

Since there is a considerable number of documents and news articles can be relatively long, the number of vertices and edges is large. For this reason, and due

to the fact that our goal is to identify the most relevant relationships, we selected only the top 200 edges with the highest weights. Finally, we calculated the maximum spanning tree out of the remaining graph, generating a graph that depicts the most relevant relationships in the format of a tree.

The final networks for the news written in English are presented in Figure 3 and clearly show some clusters that generally represent different themes or specific events. Some of the most relevant ones are: the "data" clusters, related to privacy, regulation and data protection, containing words like "privacy" and the name of information technology companies; the "encryption" clusters, related to the discussion towards WhatsApp's end-to-end encryption and containing words like "security" and "message"; and the "crime" clusters, with words like "police", "attack" and "arrested".

The network regarding Brazilian news articles, presented in Figure 4, also shows two of the aforementioned clusters: the "data" cluster ("dados", "usuários", "mensagens") and the "crime" cluster ("polícia", "civil"). Besides that, it presents at least two other particularly interesting clusters. The first one is related to the government blocking WhatsApp in Brazil, with words like "bloqueio", "justiça" and "operadoras"; and the second one is the "truck drivers' strike" cluster, represented by the words "caminhoneiros", "greve" and "governo".

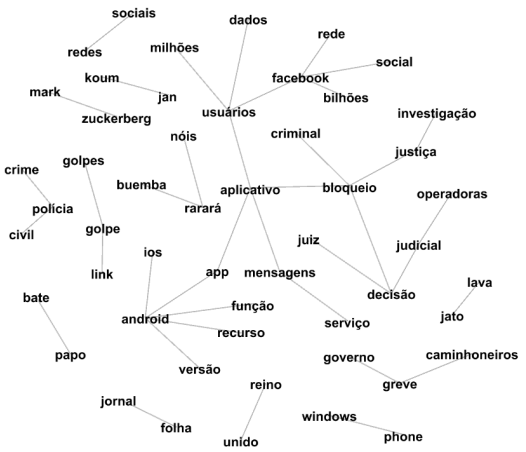


Figure 4: Co-ccurrence network for Brazilian news articles

4.5 Topics addressed

In addition to investigate the vocabulary present in news articles mentioning WhatsApp, it is also possible to find the main topics addressed in the texts included in our datasets. We used *latent Dirichlet allocation* (LDA) [BNJ03] to automatically discover topics discussed in texts. For this task, we first lowercased and tokenized all the words in the datasets. Then, we removed stop words using, once again, the lists provided by the Natural Language Toolkit (after having added the word "whatsapp" to the lists, since it appears in all texts). Finally, we ran the LDA algorithm using the Python library `spaCy`[7] for topic modeling. We used topic coherence score [NLGB10] to choose the optimum number of topics k to be returned by the algorithm. For each region and year, the LDA model returned these k topics containing terms ordered by importance in the corresponding text. We then selected the most important topic as the representative of each region and year.

Table 5 shows the top-ranked ten terms produced by our LDA model representing the main topic for each region in each year. Here, for the Brazilian articles, we translated the terms from Portuguese to English.

It is interesting to observe that, between years 2010 and 2013, the main topic in almost all regions was related to WhatsApp features, device compatibility and differences between this application and other technologies, like SMS. In the Indian subcontinent, however, the main topic of 2013 was about riots and politics.

In the Americas, in years 2016-2017, the main topics of the news articles were also related to WhatsApp features. However, in 2014 and 2015 we can observe words like "refugee", "libya" and "jihadist", probably associated with events in the Arab world. In 2018, the main topic is related to the royal British wedding.

In Brazil, in 2014, we observe a topic shift to news related to criminal scams in WhatsApp. It is interesting to note that the main topic of 2015 is related to a Brazilian court decision to block WhatsApp in the whole country (because the company did not cooperate in a criminal investigation). In 2016, year of the impeachment of president Dilma Rousseff, the main topic contains words like "dilma", "impeachment" and "lula", while the main topic in 2017 is also about politics, but containing more generic terms, such as "politics" and "government". In 2018, however, we observe a clear dominance of terms related to Brazil truck drivers' strike, considered the biggest strike in the history of the country [Phi18]. In this occasion, WhatsApp played an essential role in the organization of the strike, differently from previous protests that were mostly coordinated through Facebook and Twitter. This result reinforces the claims that WhatsApp is a valuable tool to communicate and also to share political ideas in Brazil.

In the Indian subcontinent, we note that, between

[7] `https://spacy.io/`

Table 5: Main topics for each year in each region

Year	The Americas	Southeast Asia	British Isles	Africa	Oceania	Indian subcontinent	Brazil
2010	kik, service, federal, user, message, say, livingston, rim, blackberry, growth, contact	land, asli, right, community, customary, indigenous, malaysian, government, recognition	—	—	app, phone, free, message, text, iphone, major, blackberry, service, unlimited	nokia, download, top, smartphone, mobile, skype, america, india, ovi, store	—
2011	business, small, wilton, owner, blackberry,patricio, product, plan, entepreneur, service	free, text, call, app, viber, iphone, network, phone, charge, service	charge, dutch, net_neutrality, extra, mobile, internet, law, issue, state, kpn, skype	mxit, knott_craig, market, platform, cheap, attention, buy, facebok, mobile, smartphone	skype, iphone, call, viber, free, phone, mobile, platform, android, text	message, phone, handset, android, service, blackberry, iphone, symbiam, text, app	—
2012	hutterite, wipf, colony, website, waldner,help, people, life, medium, social	hong_kong, customer, data, free, service, messaging, roam, hutchison, application, lead	app, year, game, iphone, apple communication, social, lync, facebook, network	mobile, app, call, nigerian, telecom, constitution, growth, industry, service, call, send	world, system, stop, late, malware, ransomware, sonicwall, prevent,learn wannacry	phone, nokia, asha, service, price, launch, company, internet, cost, news	client, message, app, reconquer, telefonica, europe, launch, viber, intensify, instantaneous
2013	blackberry, bbm, user, company, device, playbook, service, app, popularity, release	lau, speak, mobile, tencent, commercial,chat, application, martin, ltd, conference	screen, small, blackberry, feel, keyboard, camera, application, distraction, fibre, carbon, design	science, kelemu, agricultural, research, african, woman, school, international, develop, award	privacy, user, woman, data, message, dutch, policy, server, agency, address_book	communal, india, people, indian, blood, muslim, riot, politician, secular, muzaffarnagar	market, technology, brazil, china, consume, difficulty, emerging, attempt, domesticate, invade
2014	canada, refugee, game, family, libya, trip, help, team, furniture, huddle	government, right, party, state, political, law, country, leader, election	party, johnson, cohen, birmingham, former, leader, britain, secretary, prime_minister, vote, election	show, music, event, african, art host, competition, world, winner, black	minister, party, leader, government, prime_minister, election, national, senator, republican, president	pakistan, indian, terrorist, attack, muslim kill, freedom, kashmir, army, journalist	criminal, victim, page, victim, false, security, click, virus, browser, federal
2015	jihadist, cabinet, edward, plausible, outrage, nude, csec, guido, chamber, trove, inherently	social_media, job, linkedin, online, twitter, facebook, company, professional, jobseeker, network	terrorist, security, attack, intelligence, nisman, law, communication, cameron, gchq, encryption	burundi, man, white, election, protest, nkurunziza, president, police, party, bujumbura	refugee, camp, boat, data, web, australian, service, phone, use, communication	geeta, pakistan, woman, girl, ansar_burney, karachi, sushma_swaraj, police, comissioner, raghavan	carriers, application, business, service, block, justice, brazil, decision, voice, president
2016	business, canada, chera, border, cbsa, trade, red_tape, cfib, government, agency, raise	hong_kong, china, chow, market, hktdc, president, wechat, product, fair, party	scotland_yard, religious, muslim, country, british, authority, murder, cafe, pope, bbc	nakuru, group, political, youth, party, member, jubilee, nyamira, governor, leadership	refugee, syrian, aleppo, alkhuder, homescreen, syria, earthquake, greece, aircraft, assad	police, karachi, punjab, ranger, arrest, pakistan, kashmir, kill, medium, protest	government, year, president, fear, lula, dilma, work, impeachment, brazil, police
2017	market, report, company, service, user, include, help, facebook, mobile, information	group, lam, chat, chat_group, responsible, content, campaign, china, service, team	attack, school, police, masood, westminster isis, terrorist, arrest, kill, birmingham	president, nasa, kenyan, raila, leader, election, political, party, iebc, court	immigrant, country, immigration, trump, employee, visa, ban, refugee, policy, president	afghanistan, kashmir, india, taliban, trump, policy, war, obama, american, troop	police, politics, government, woman, demonstration, security, geddel, prisoner, regime, arming
2018	meghan, harry, church, england, ceremony wedding, prince, kate, vow, royal_wedding	facebook, zuckerberg, data, scandal, mistake, platform, prevent, ad, authority, issue	facebook, people, world, internet, company, technology, social_media, online, change, new	student, school, university, high, education, parent, water, health, disease, study	national, party, government, leader, state, support, michael, australian, election, member	china, wechat, tencent, newsguard, app, chinese, company, traffic, mall_road, authority	government, president, truck_drivers, strike, minister, military_coup, group, support, deputy, world_cup

years 2013 and 2017, the main topics were related to political themes. In the year 2013, for example, words like "riot", "muslim" and "muzaffarnagar" are associated with the riots in Muzaffarnagar, when some rioters used WhatsApp to promote violence. In the years 2014-2016, rumors on terrorist attacks were disseminated through WhatsApp. In 2017, the main topic seems to be associated to the decision of US president Donald Trump to not withdraw its troops from Afghanistan.

In Africa, in 2014, the words "burundi", "election" and "protest" are related to protests that occurred during the Burundian election. In this occasion, the government temporarily blocked messaging services, including Facebook, WhatsApp and Twitter [Vir16]. In 2017, words like "election" and "president" are associated with the suspicion that disinformation and fake news were being used to influence Kenyans during the elections [Sam17].

There is also a clear dominance of words associated with terrorist attacks in 2015 and 2017 in the British Isles. These words are related to the use of

WhatsApp to organize these acts [BBC17]. In Southeast Asia, news on WhatsApp are generally associated with comparisons with WeChat and, in the year 2018, news in this region were associated with the Facebook–Cambridge Analytica data scandal. In Oceania, in the years 2015-2017, the main topics were associated with refugees and immigration. WhatsApp played an important role during the Syrian Civil War in these years, since journalists and individuals living there used WhatsApp to communicate with people of foreign countries [Boh17].

These results show that WhatsApp usage is highly associated with important political events in several regions of the world – particularly in Africa, Brazil and India. The shift in the main topics addressed in the regions before 2013 (that were related to WhatsApp features and device compatibility) to, in the following years, political and criminal themes confirms results (presented in previous sections) that indicate a gradual increase in the association of this application with social and political situations.

4.6 Polarity

Our final investigation sheds light in another dimension of the news articles containing the term "whatsapp": now, we analyze the *polarities* of the articles – that is, whether the expressed opinions in the texts are mostly positive, negative or neutral. Here, we are interested in analyzing how the polarity of news articles related to WhatsApp changes over time and in different regions.

To do this, we performed sentiment analysis in each of the articles in our datasets using `SentiStrength` [TBP+10], a tool that estimates the strength of positive and negative polarities in texts. This tool receives as input pieces of text and returns a score that varies from -4 (negative) to +4 (positive).

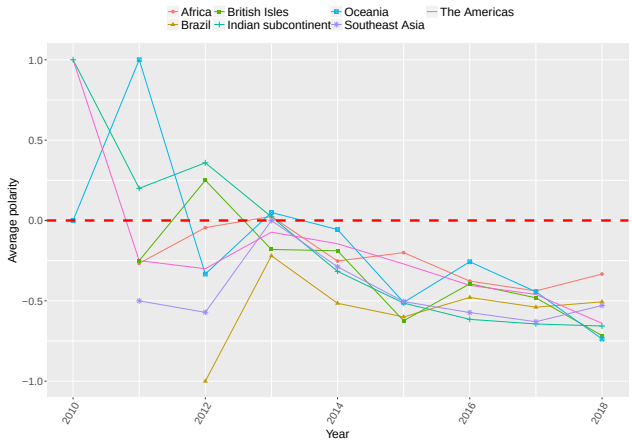


Figure 5: Average polarity of news articles from different regions containing the term "whatsapp" over time

Figure 5 depicts the average polarity of the news articles in each region and in each year, both in NOW Corpus and in the dataset of Brazilian articles. We observe a major dominance of negative polarities in almost all regions and years, but especially after 2013. News articles containing the term "whatsapp" are becoming more negative over time probably because of the nature of the news articles themselves: in Africa, for instance, the term "whatsapp" occasionally appeared in news articles about refugees[8]; in India, in articles about the spread of fake news that resulted in violence[9]; in Southeast Asia and the Americas, in news about the promotion of violence[10]; in Brazil, in news concerning criminal scams[11].

4.7 Summary of results

The most relevant findings presented in this section can be summarized as follows:

- the interest for the term "whatsapp" is constantly increasing over the years, as indicated by the rise of news about this tool and of Google Search queries for this term (Section 4.1);
- this interest is being accompanied by a change of framing around the term "whatsapp" in the media – from topics regarding WhatsApp features and technology to those related to misinformation, politics and criminal scams (Sections 4.1, 4.3, 4.4, 4.5);
- the polarity of news articles containing the term "whatsapp" is becoming more negative over time, probably due to the fact that this tool is being gradually more associated with crimes, violence and fake news (Section 4.6).

5 Concluding Remarks

In this paper, we present a quantitative analysis on the public perception of the messaging tool WhatsApp in news articles. For conducting our analyses, we used two datasets that cover the whole history of the application since its release for Android devices in 2010 until May 2018. The first of these datasets is a corpus of news articles written in English and published from 2010 to 2018 in 20 countries, while the second one contains Brazilian news articles published from 2012 to 2018. We also used data collected from Google Trends in one of our analyses.

Here, we investigated how media sources from different parts of the world have been reporting stories related to WhatsApp and whether the rise of the public

[8] `https://bit.ly/2GSTlo7`
[9] `https://bit.ly/30qdcmn`
[10] `https://nyti.ms/2EehjIJ`
[11] `https://bit.ly/2VNnoGY`

interest in this application over time was accompanied by changes on its perception by the media. We observed changes in the vocabulary, in the mentioned entities, in the addressed topics and in the polarity of the articles mentioning the tool WhatsApp in our datasets. In particular, we noticed a shift on media perception in almost all analyzed regions from the period before 2013 – when the focus was on WhatsApp features and device compatibility – to the following years – when the application started to be gradually more associated with misinformation, manipulation and extremism, as well as with political and criminal activities.

The techniques and approaches proposed here can be used to measure the media perception of any company (or entity in general), but WhatsApp was chosen due to its influence in information (and misinformation) dispersion and to the fact that it has been related to topics such as extremism, corruption and political propaganda. In future works, we intend to add more analyses, use news articles from others regions of the world where WhatsApp is popular (e.g. Germany, Indonesia, Malaysia) and compare the perception of WhatsApp in the media with the perception of it in other sources, like social networks and news articles comments. Also, we plan to compare the public perception of WhatsApp with the one of similar tools (e.g. Telegram, Facebook Messenger, WeChat) in order to understand which of them are more likely to be mentioned in certain types of news – for instance, in political or crime-related news.

6 Acknowledgements

This work was partially supported by CNPq, CAPES, FAPEMIG and the projects InWeb, MASWEB and INCT-Cyber.

References

[ALGB13] Alberto Acerbi, Vasileios Lampos, Philip Garnett, and R Alexander Bentley. The expression of emotions in 20th century books. *PLOS ONE*, 8(3):e59030, 2013.

[BBC17] BBC News. WhatsApp must not be 'place for terrorists to hide'. Retrieved from `https://bbc.in/2ZIHKz8`. Accessed on May 16, 2019, 2017.

[BLK09] Steven Bird, Edward Loper, and Ewan Klein. *Natural language processing with Python*. O'Reilly Media Inc., 2009.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[Boh17] Lauren Bohn. Syrian history is unfolding on WhatsApp. Retrieved from `https://bit.ly/2Va1VaU`. Accessed on May 16, 2019, 2017.

[Cam13] César Nardelli Cambraia. Da lexicologia social a uma lexicologia sócio-histórica: caminhos possíveis. *Revista de Estudos da Linguagem*, 21(1):157–188, 2013.

[CdO13] Karen Church and Rodrigo de Oliveira. What's up with whatsapp?: Comparing mobile instant messaging behaviors with traditional sms. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '13, pages 352–361, New York, NY, USA, 2013. ACM.

[CMC+18] Evandro Cunha, Gabriel Magno, Josemar Caetano, Douglas Teixeira, and Virgilio Almeida. Fake news as we feel it: perception and conceptualization of the term "fake news" in the media. In *Proceedings of the 10th International Conference on Social Informatics (SocInfo 2018)*, 2018.

[CMG+14] Evandro Cunha, Gabriel Magno, Marcos André Gonçalves, César Cambraia, and Virgilio Almeida. How you post is who you are: Characterizing Google+ status updates across social groups. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media (HT'14)*, pages 212–217, New York, NY, USA, September 2014. Association for Computing Machinery (ACM).

[Con18] John Constine. WhatsApp hits 1.5 billion monthly users. $19b? not so bad. Retrieved from `https://tcrn.ch/2LdlavD`. Accessed on May 16, 2019., 2018.

[Dav13] Mark Davies. Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day. Available online at `https://corpus.byu.edu/now/`, 2013.

[FALW+13] Ilias Flaounas, Omar Ali, Thomas Lansdall-Welfare, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. Research methods in the age of digital journalism: Massive-scale automated analysis of news-content – topics, style and gender. *Digital Journalism*, 1(1):102–116, 2013.

[FCB16] Ethan Fast, Binbin Chen, and Michael S. Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4647–4657, New York, NY, USA, 2016. ACM.

[FCSD15] P. Fiadino, P. Casas, M. Schiavone, and A. D'Alconzo. Online social networks anatomy: On the analysis of facebook and whatsapp in cellular networks. In *2015 IFIP Networking Conference (IFIP Networking)*, pages 1–9, May 2015.

[FTA+10] Ilias Flaounas, Marco Turchi, Omar Ali, Nick Fyson, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. The structure of the EU mediasphere. *PLOS ONE*, 5(12):e14243, 2010.

[GB11] Kristina Gulordava and Marco Baroni. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71. Association for Computational Linguistics, 2011.

[Goe18] Vindu Goel. In India, Facebook's WhatsApp plays central role in elections. Retrieved from `https://nyti.ms/2Il6uV3`. Accessed on May 16, 2019, 2018.

[GT18] Kiran Garimella and Gareth Tyson. Whatsapp, doc? A first look at whatsapp public group data. *CoRR*, abs/1804.01473, 2018.

[GWCG18] Gaoyang Guo, Chaokun Wang, Jun Chen, and Pengcheng Ge. Who is answering to whom? finding "reply-to" relations in group chats with long short-term memory networks. In Wookey Lee, Wonik Choi, Sungwon Jung, and Min Song, editors, *Proceedings of the 7th International Conference on Emerging Databases*, pages 161–171, Singapore, 2018. Springer Singapore.

[Lee11] Kalev Leetaru. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16(9), 2011.

[LWSVC14] Thomas Lansdall-Welfare, Saatviga Sudhahar, Giuseppe A Veltri, and Nello Cristianini. On the coverage of science in the media: A big data study on the impact of the Fukushima disaster. In *2014 IEEE International Conference on Big Data*, pages 60–66. IEEE, 2014.

[Mat53] Georges Matoré. *La méthode en lexicologie: domaine français*. Didier, Paris, 1953.

[MGB17] Andres Moreno, Philip Garrison, and Karthik Bhat. Whatsapp for monitoring and response during critical events: Aggie in the ghana 2016 election. In *14th Int. Conf. on Information Systems for Crisis Response and Management*, 2017.

[MSA+11] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[NFK+18] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, David AL Levy, and Rasmus Kleis Nielsen. Reuters institute digital news report 2018. `http://www.digitalnewsreport.org/`. Accessed on May 4, 2018, 2018.

[NLGB10] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 100–108, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Phi18] Dom Phillips. Truckers' strike highlights 'a dangerous moment' for Brazil's democracy. Retrieved from `https://bit.ly/2HlmQMm`. Accessed on May 16, 2019., 2018.

[PTHS12] Alexander M Petersen, Joel Tenenbaum, Shlomo Havlin, and H Eugene Stanley. Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports*, 2, 2012.

[Rot14] Steffen Roth. Fashionable functions: A Google Ngram view of trends in functional differentiation (1800-2000). *International Journal of Technology and Human Interaction*, 10(2):34–58, 2014.

[RSS+18] Avi Rosenfeld, Sigal Sina, David Sarne, Or Avidov, and Sarit Kraus. A study of whatsapp usage patterns and prediction models without message content. *CoRR*, abs/1802.03393, 2018.

[RU11] Anand Rajaraman and Jeffrey David Ullman. *Data Mining*, page 1–17. Cambridge University Press, 2011.

[Sam17] Nanjira Sambuli. How Kenya became the latest victim of 'fake news'. Retrieved from `https://bit.ly/2XUS5GM`. Accessed on May 16, 2019, 2017.

[SHS+16] Michael Seufert, Tobias Hoßfeld, Anika Schwind, Valentin Burger, and Phuoc Tran-Gia. Group-based communication in whatsapp. In *IFIP Networking Conf. (IFIP Networking) and Workshops, 2016*, pages 536–541. IEEE, 2016.

[Sta18] Statista. Share of population in selected countries who are active WhatsApp users as of 3rd quarter 2017. Retrieved from `https://bit.ly/2k9ZV0y`. Accessed on May 16, 2019, 2018.

[TBP+10] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558, December 2010.

[Vir16] Thierry Vircoulon. Burundi turns to WhatsApp as political turmoil brings media blackout. Retrieved from `https://bit.ly/1U6m7OS`. Accessed on May 16, 2019, 2016.

[Wat18] Jim Waterson. Fears mount over WhatsApp's role in spreading fake news. Retrieved from `https://bit.ly/2MzEHD6`. Accessed on May 16, 2019, 2018.